REVIEW

Bridging proteomics and systems biology: What are the roads to be traveled?

Serhiy Souchelnytskyi

Ludwig Institute for Cancer Research, Uppsala, Sweden

The comprehensive study of proteomes has become an important part of attempts to uncover the systemic properties of biological systems. Proteomics provides data of a quality which increasingly fulfills strict requirements of systems biology for quantitative and qualitative information. Notably, proteomics can generate rich datasets that describe dynamic changes of proteomes. On the other hand, large-scale modeling requires the development of mathematic tools that are adequate for the processing of largely uncertain biological data. In this review, recent developments that pave the way for the integration of proteomics into systems biology are discussed. These developments include the standardization of data acquisition and presentation, the increased comprehensiveness of proteomics studies in description of functional status, localization and dynamics of proteins, and advanced modeling approaches.

Received: March 4, 2005
Revised: May 16, 2005
Accepted: May 18, 2005

Keyword:
Systems biology

1 Proteomics and systems biology

1.1 Proteomics and modeling of biological processes

The dynamic complexity of biological processes has been less well understood, when compared to many physical and chemical processes. Apparently, sending a man to the Moon is less complicated than the full understanding of how bruises heal. This situation is about to change: the sequencing of a number of genomes, large-scale explorations of transcriptomes, proteomes and metabolomes, and a huge volume of directed studies inspire hope that we will be able to describe a living creature in the strict language of mathematics. Most importantly, there is a hope that we will be able to design better treatments and predict outcomes for human diseases. The development of modeling tools fuels these expectations, with the dawn of systems biology. Study of the systemic properties of biological systems, as systems biology can be defined, has already provided successful examples, *e.g.*, insights into the physiology of the heart, diabetes, asthma and cancer (reviewed in [1, 2]).

Building and analysis of models of biological processes comprises a number of modeling tools, and addresses biological complexity on the levels from biochemical reactions and cell physiology to behavior and evolution [3, 4]. Here and through-out this review the term "model" refers to description of biological processes in mathematical terms, without discrimination of mathematical tools. Consequently, modeling is defined as "the application of methods to analyze complex, real-world problems to make predictions about what might happen with various actions" (see Computational Science Glossary, wofford.info/ecs/glossary/terms.htm; Table 1). Modeling tools cover a broad range of mathematical methods, from systems of differential equations to statistical correlation tools [3, 4]. Some of the tools require detailed knowledge about components, *e.g.*, to build a systems of differential equations for modeling of a signaling

Correspondence: Serhiy Souchelnytskyi, Ludwig Institute for Cancer Research, Box 595, BMC, 751 24, Uppsala, Sweden
E-mail: serhiy.souchelnytskyi@licr.uu.se
Fax: +46-18-16-0420

Abbreviations: ABPP, activity-based proteome profiling; **BEMAD**, β-elimination followed by Michael addition with DTT; **FAK**, focal adhesion kinase; **FRET**, fluorescence resonance energy transfer; **GFP**, green fluorescence protein; **ICPL**, isotope-coded protein label; **MAPK**, mitogen-activated protein kinase; **PEDRo**, proteome experiment data repository; **SBML**, systems biology markup language; **TGF**, transforming growth factor-β; **XML**, extensible markup language; **YFP**, yellow fluorescence protein

Table 1. Glossary of terms

Protein species	Proteins with a particular set of PTMs. For instance, phosphorylated and non-phosphorylated proteins will define two separate, though relative species.
Connection	Indicates a physical or functional interaction between two protein species. Here it is used as an equivalent to edges and strings.
Dependencies	Definition of the influence of one protein species on another.
Variables	Components of a system, *e.g.*, protein species.
Model of a biological process and modeling	A model can be defined as a description of biological processes in mathematical terms. Modeling can be defined as the application of methods to analyze complex problems to make predictions about what might happen with various actions (wofford.info/ecs/glossary/terms.htm). Models can be described using various tools, *e.g.*, simple graph representation of dependencies, systems of differential equations, Bayesian and Boolean networks, statistical correlation tools, etc. Input data provide information with different degree of details about protein species, *e.g.*, about concentration, localization, functional status, dependencies and dynamics. High level of details is required for high-definition modeling, which is often based on use of differential equations. Less detailed datasets can be used for low-definition modeling which requires description of relation between protein species, and does not require information about absolute quantity and functional status of proteins.
Markup languages XML and SBML	eXtensible Markup Language (XML) is a general markup language to structure data. Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biochemical reaction networks (http://sbml.org). SBML allows recording of information about biological processes in a format which can be used in modeling of these processes. SBML also allows exchange of biological data between various modeling software.
Polypeptide growth factors TNFα, EGF, PDGF, FGF, TGFβ, BMP and their signaling	Polypeptide growth factors are potent regulators of cell proliferation, migration, differentiation and death. They are secreted molecules, and bind to their specific transmembrane receptors on the cell surface. This binding activates receptor-associated kinases, which phosphorylate specific substrates. Phosphorylated substrates initiate cellular responses to the growth factors by binding to other proteins or changing their enzymatic activities.

pathway, knowledge of concentration of components and kinetic parameters of reactions in this pathway is required. Data with less precise information can be analyzed by other tools, *e.g.*, to build a model based on a Bayesian network, it is sufficient to know relations between studied components of a model. However, common for all modeling tools is the requirement of information about quantity and identity of components of a model, and knowledge of dependencies and dynamics of relation between these components (Figs. 1, 2). As proteins are key molecules of any living organism, they are essential components in modeling of biological processes. In-depth studies of selected proteins provide valuable information about features of these proteins, but such studies seldom explore proteins in numbers that would be significant for modeling. Therefore, a comprehensive and simultaneous analysis of hundreds of proteins is required for generation of datasets suitable for modeling.

Proteomics is a large-scale technology which provides a global overview of proteomes (reviewed in [5, 6]). Information about protein expression, rates of synthesis and degradation, PTMs, enzymatic activities, structure, localization and interacting partners can be generated by modern proteomics techniques, although with various degree of preciseness. The comprehensiveness and increasing confidence in proteomics data make them suitable for large-scale modeling [1, 5, 6–8]. Tools for acquisition of biological data in a format that can be used for modeling have been under development. Systems Biology Workbench represents one of such initiatives, with XML-based Systems Biology Markup Language (SBML) being an important step in development of common standards for description of biological datasets, *e.g.*, proteomics, genomics, metabolomics and in-depth studies [9].

Recent progress in proteomics has provided techniques for detection and identification of proteins in (semi)-quantitative ways; these techniques have been discussed [5, 6]. Selected signaling pathways have also been reviewed, including studies of PTMs and protein activities, interactions and localization [10–12]. However, an analysis of proteomics data integration in modeling efforts requires an overview of proteomics technologies from the prospective of their suitability to deliver information that can be used for modeling. There is also a need to evaluate how modeling tools can adapt to specifics of proteomics data. An informative modeling has minimal requirements on the input data, as mentioned above, and proteomics has its limitations in description of proteomes. The question arises concerning how proteomics fulfills modeling requirements. Here five proteomics-related issues of importance for the integration of proteomics and systems biology are discussed: (1) presentation of data, (2) quantification of proteins, (3) the functional status of proteins (activity and interactions), (4) localization, and (5) the dynamics of proteomes (Fig. 2).

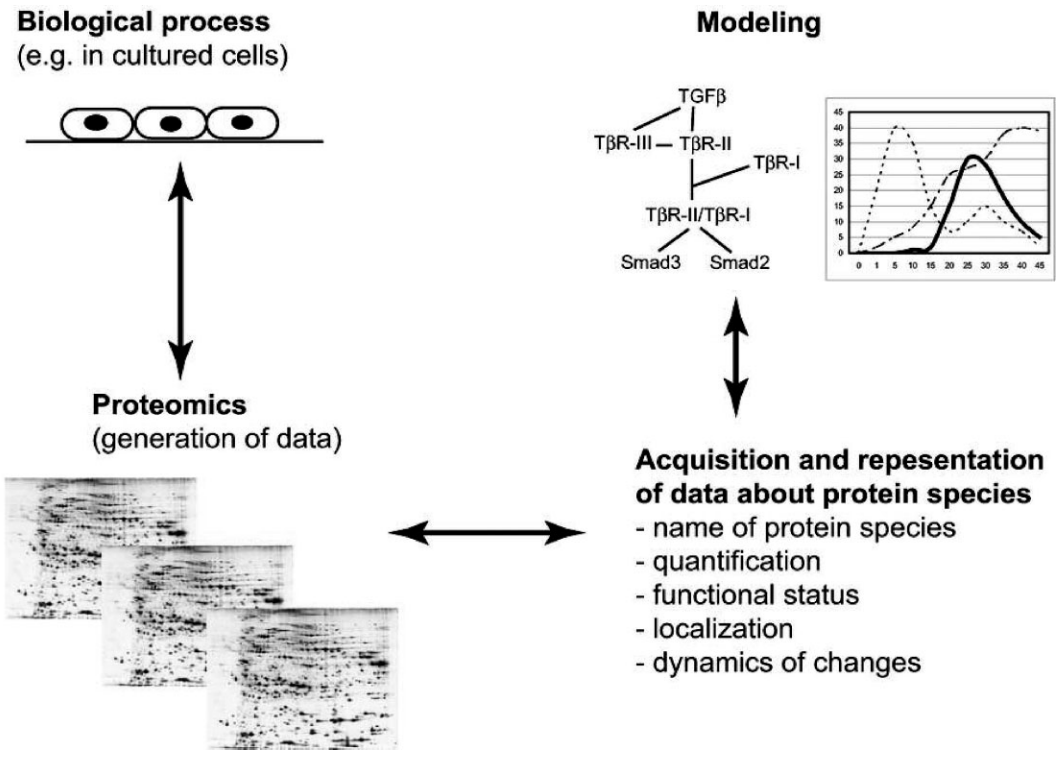

Figure 1. Steps of biological modeling using proteomics data. In this figure, experimental models include any biologically relevant processes in living systems, and are represented by cultured cells as an example. Proteomics includes the variety of technologies to study proteins expression, functional activities, PTMs, protein-protein interactions, localization and dynamics. Proteomics is exemplified by 2-DE. Proteomics datasets have to be acquired and represented in formats that are compatible with modeling tools. It is represented by listing information which is required for presentation of biological data. Proteomics datasets which are translated into modeling-compatible formats can be then used for modeling. Here, modeling is represented by a relation network between TGFβ, its receptors (TβR-I, TβR-II, TβR-III) and receptor substrates (Smad2 and Smad3), and a graph which illustrates changes in protein concentrations. The arrows between the main steps are double-headed, as modeling tools influence requirements for data deposition and representation. Data deposition and representation also affect design of proteomics experiments, and proteomics technologies may affects selection of a biological model.

1.2 Presentation of data

What is the best format for collecting information about proteins? This information has to provide a unique identifier for a protein, *e.g.*, a unique name or tag, and to describe the features of a protein, *e.g.*, molecular mass, PTMs, and its functional status. It has also to describe technical details of how the information was generated, and to be in a format which is compatible with modeling tools. Development of XML-supporting formats for the representation of proteomics data is an important step towards addressing these requirements. PEDRo, PSI-MI, mzXML, AGML and similar initiatives have laid the groundwork for the systematic recording of proteomics experiments, and have been described elsewhere [13–16]. Repository databases that are based on these formats allow documentation of protein identities and technical details of experiments. XML-supported standardization and exchangeability of proteomics data is the basis for an open proteomics database in which proteomics experiments can be freely accessible for multiple applications. Unified representation in XML format also simplifies the importation of proteomics data into modeling tools. XML-based SBML is the basis for representation of biological models, and it allows exchange of data between different models in a well-defined format (Fig. 2). The unified representation requires adoption of a nomenclature for protein species and their presentation in functional states, *e.g.*, PTMs status and interactions with other proteins. The Gene Ontology project provides nomenclature [17], which can be the basis for nomenclature of protein species, as Gene Ontology-adopted protein names can be core-names for the description of modified proteins and proteins in complexes, *e.g.*, protein species.

Repository databases provide the framework for the description of static features of proteins, while tools for the systematic description of protein dynamics are less developed. The last requires connecting protein species with defined PTMs in defined complexes and in a defined loca-

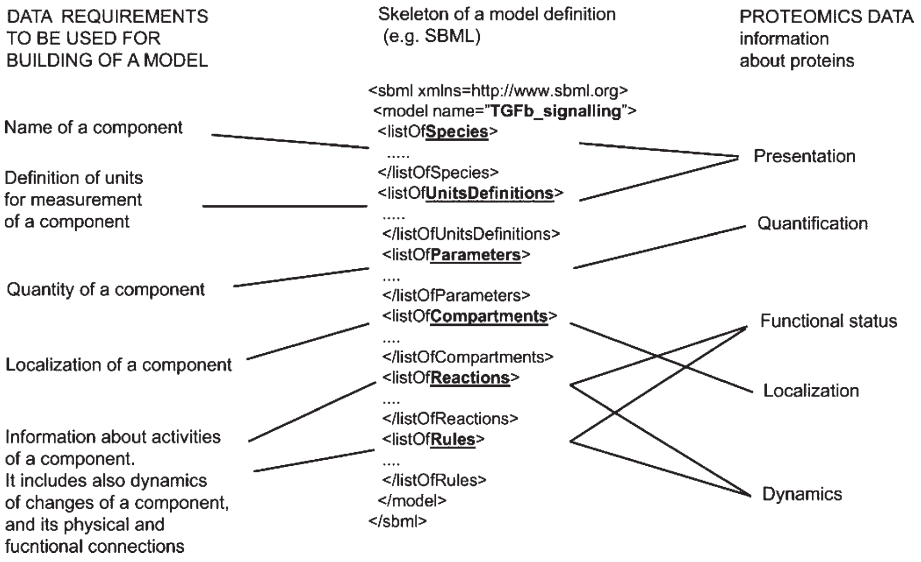

Figure 2. Proteomics data in a context of requirements of modeling tools. Examples of the type of information about studied components required for a modeling are shown in the left part of the figure, as requirements to data to be used in modeling. Types of information about proteins that can be generated by a proteomics experiment are shown in the right part of the figure, as presentation, quantification, functional status, localization and dynamics. In the center, a skeleton of the SBML scheme for acquisition and presentation of biological data is shown in a format compatible with modeling tools [9].

tion, through a time interval. This task can be solved in a framework of modeling formats, similar to formats supporting SBML. These formats incorporate description of the functional, spatial and quantitative features of proteins (see www.sbml.org, for examples) [9]. It is important to note that the SBML and related efforts allow the building up and expansion of an existing model upon generation of new data. Thus, presentation of proteins, as the first issue in integration of proteomics and systems biology, is being addressed by the creation of repository databases. However, definition of protein species in these databases, especially proteins with PTMs, is still a challenge which has to be solved in the nearest future.

1.3 Quantification of proteins

Quantitative presentation of protein species is essential for informative modeling. 2-DE and LC-protein/peptide MS-based techniques are the most common approaches to unbiased evaluation of protein expression. These techniques have limited coverage of proteomes, which is estimated to be 10–20% of a total number of proteins in cells [5, 6]. Higher coverage of open reading frame (ORF)-defined proteins has been claimed with use of peptide/MS-based proteomics, as compared to 2-DE [5]. However, combined 2-D gels generated with various narrow pH range gels provide much higher coverage, as compared to a single 2-D gel. In addition, 2-D gels provide information about modifications of proteins, *e.g.*, proteolytic cleavage, phosphorylated forms, etc. [6].

Protein enrichment techniques may increase coverage of selected protein groups, but they decrease the comprehensiveness of proteome analysis. As an example, antibodies, IMAC and lectins allow enrichment for phosphorylated or glycosylated proteins [18, 19]. Chemical modification of phosphorylated or glycosylated amino acid residues is another way to enrich proteins with these PTMs, and have been described in recent publications [20–22]. However, with such enrichment techniques, non-modified proteins will be lost for analysis.

Protein chips have been developed to enrich and measure expression of selected proteins [19]. Various protein-capturing agents have been described, *e.g.*, antibodies, aptamers, and proteins or their domains [19]. Protein chips are useful for monitoring selected proteins, but novel non-annotated proteins would not be detected. Protein chips are known as expression, interaction and structural arrays [18, 19]. Chips are important for the interrogation of any models which analyze protein expression, protein-protein interactions and protein modifications. Protein chips allow fast and reliable analysis of a relatively large number of components, which is crucial for such a model under interrogation.

2-D gel-based proteomics allow quantification by staining of protein spots. Fluorescent probes, CBB or silver staining are the techniques used most often; they give acceptable evaluation of a relative expression of proteins (reviewed in [6]). Staining allows generally good, although variable, linearity and sensitivity of protein detection. Modified labeling of proteins with DIGE allows evaluation of relative levels of proteins in samples that are mixed after labeling with different fluorescent dyes, and run in one gel. Thus, DIGE requires running of fewer gels, as compared to techniques with a single type of staining. Staining is an efficient way to visualize proteins, but it provides assessment of a relative, and not an absolute protein concentration.

The relative expression of proteins can be estimated by comparison of selected peptides using isotope-coded affinity tag (ICAT) and stable isotope labeling with amino acids in cell culture (SILAC and ICPL) techniques [23–26]. The combined fractional diagonal chromatography (COFRADIC) approach can also be used for evaluation of protein expression [27]. However, peptide/MS-based proteomics approximates iden-

tification of few peptides to identification of a full-length protein. This introduces uncertainty for protein identification, and inability to evaluate levels of protein species with a defined spectrum of PTMs. The same peptide may originate from proteins which are modified on a site that may be important for its biological activity, as well as from proteins which are non-modified at this site and are inactive. Moreover, peptide/MS-based proteomics is not suitable for the evaluation of absolute quantities of proteins, although it gives a good estimation of the levels of modification at a selected amino acid residue [5, 6, 23, 28, 29]. Thus, peptide/MS-based proteomics can be an acceptable method of evaluating the relative expression of proteins under the condition that the expression levels of full-length proteins are confirmed.

An interesting possibility was proposed to estimate the absolute quantity of proteins by metabolic labeling of cells with ^{35}S-labeled amino acids to saturation [30, 31]. The quantity of proteins would be calculated from the absolute radioactivity incorporated in the proteins, and corrected to the level of specific radioactivity of a probe and a number of cysteine and methionine residues in a protein. This technique also allows the evaluation of synthesis and turnover rates of cellular proteins [30, 31].

Thus, quantification of relative levels of protein expression, including relative levels of PTMs, can be achieved in a number of ways (Fig. 3). If quantification data are to be used for modeling, it is important that relative expression levels can be compared not only between various paired samples, but also normalized to levels of reference "housekeeping" proteins within the same sample. Further measurements of the absolute concentration of "housekeeping" proteins allow calculation of absolute quantities of the proteins of interest, which is the most suitable for a high-definition modeling, *e.g.*, for modeling which requires relatively detailed information about proteins and their dynamics. High-definition models are often based on use of systems of differential equations, and require knowledge of protein concentration and kinetic parameters [4]. In many proteomics studies only relative expression levels of proteins can be measured. These less-detailed data, *e.g.*, relative levels of protein expression, may also be informative for description of dependencies between proteins using low-definition modeling, *e.g.*, modeling which requires description of relation between proteins, and does not require detailed information about quantity or functional status of proteins [4]. An evaluation of relative levels of protein expression is well developed in modern proteomics, and it allows building of low-definition models. However, an acquisition of absolute concentrations of proteins for high-definition models requires further developments of proteomics techniques.

1.4 Functional status of proteins

The functional status of proteins is the most difficult parameter to evaluate and describe in a comprehensive way. The functional potential of a protein is defined by its PTMs and by interacting molecules. Functional status describes features of a protein that will define how this protein will affect a modeled system. Approximately 200 reversible and non-reversible covalent PTMs create an astonishing variability of protein species [3, 10, 32]. Comprehensive simultaneous analysis of all PTMs in a proteome is still unrealistic due to technical limitations. However, methods of studying phosphorylation, glycosylation, acetylation, ubiquitination, sumoylation and methylation have been developed. As these PTMs are involved in crucial regulatory processes in cells, there is strong interest in their study. As an example, phosphorylation on serine, threonine and tyrosine residues are essential for regulation of cell proliferation, differentiation, migration and apoptosis [11, 12, 32–34]. Initiation of signaling downstream of polypeptide growth factors such as EGF, TGFβ, PDGF, TNFα and FGF, is dependent on activation of receptor tyrosine or serine/threonine kinases. Activated kinases then phosphorylate a number of substrates that regulate cellular function [10–12, 33, 34]. PTMs also trigger protein-protein interactions, *e.g.*, interactions mediated by

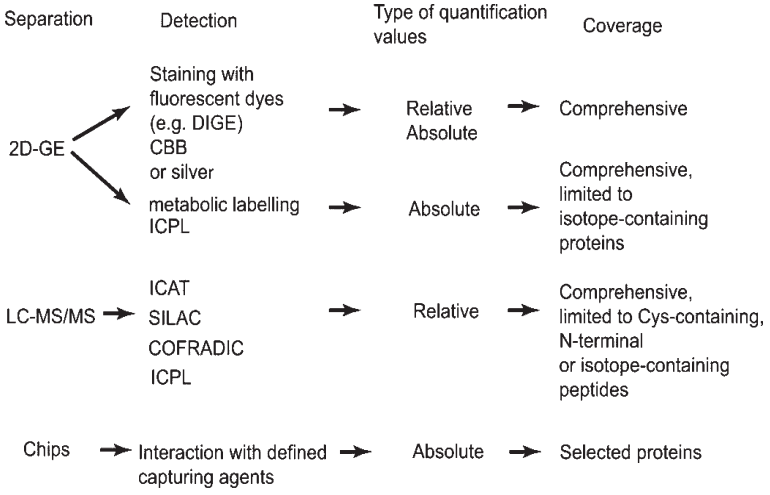

Figure 3. Quantification of proteins with various proteomics techniques. Comprehensiveness (coverage) and absolute or relative values of protein quantification (type of values) are indicated for the most commonly used proteomics separation and detection techniques.

phosphorylated tyrosine in one protein and Src-homology 2 (SH2) or phosphotyrosine binding (PTB) domains in another protein initiate a cascade of signaling events [11, 12, 32].

Modified proteins are often enriched by the use of capturing agents that recognize selected PTMs. Phosphoproteins can be enriched by immunoprecipitation with specific antibodies and by IMAC [5, 6, 24, 25, 28, 35]. Phosphorylation has also been studied by metabolic labeling of proteins with radioactive [^{32}P]phosphate, followed by 2-DE [36, 37]. Phosphorylations on serine, threonine and tyrosine residues have been studied using β-elimination-directed cleavage at the site of modification, by fluorescent dyes, and by chemical modification of phosphorylated residues [20–22, 38]. MS-based proteomics has proven to be an efficient tool for identifying modified peptides; phosphorylated peptides may be enriched with specific antibodies or by IMAC [5, 6, 23–25, 28]. Glycosylated proteins have been successfully enriched with lectins [39]. The β-elimination followed by Michael addition of DTT (BEMAD) technique allows identification of O-glycosylation sites in purified proteins [22]. Ubiquitination is an important triggering mechanism for protein degradation, and more than a thousand proteins have been identified using pull-down of His-tagged ubiquitin in *Saccharomyces cerevisiae* [40]. The analysis of SUMO-2-interacting proteins identified eight novel targets of sumoylation [41]. Thus, tools to study phosphorylation, glycosylation, ubiquitination and sumoylation have been developed (Fig. 4).

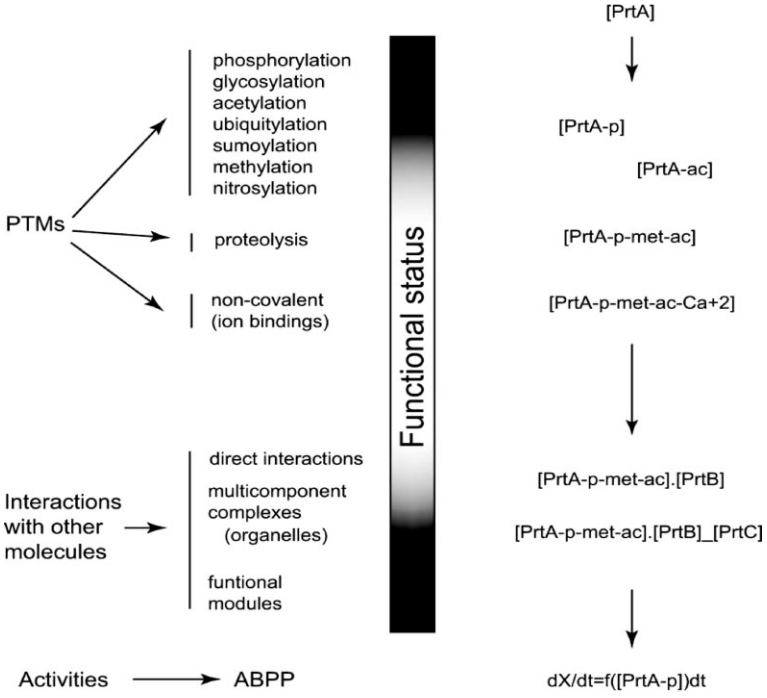

Figure 4. Functional status of proteins is defined by PTMs and interactions with other molecules. PTMs can be reversible or not, and covalent or non-covalent, as indicated. A number of PTMs are indicated, *e.g.*, phosphorylation, glycosylation, acetylation, methylation, nitrosylation, ubiquitination, sumoylation, and proteolytic cleavage. Non-covalent modifications are often by high affinity binding of ions, *e.g.*, Ca^{2+}, Mg^{2+}, etc. Interactions with other molecules can be direct or indirect. Indirect interactions can be deduced from descriptions of components of purified physical complexes, or from description of functional complexes. Comprehensive evaluation of activities of proteins can be achieved by activity-based proteome profiling (ABPP). On the right side examples of presentations of protein species as modified proteins are shown, as components of complexes, and as species with defined functional activities. [PrtA], non-modified protein A; [PrtA-p], phosphorylated protein A; [PrtA-ac], acetylated protein A; [PrtA-p-meth-ac], phosphorylated, methylated and acetylated protein A; [PrtA-p-meth-ac-Ca2+], phosphorylated, methylated and acetylated protein A in complex with Ca^{2+} ion; [PrtA-p-meth-ac].[PrtB], phosphorylated, methylated and acetylated protein A directly interacting with protein B; [PrtA-p-meth-ac].[PrtB]_[PrtC], phosphorylated, methylated and acetylated protein A directly interacting with protein B, and forming complex, but not interacting directly, with protein C; dX/dt=f([PrtA-p])dt, the definition of a functional activity of phosphorylated protein A.

Acetylation and methylation are two PTMs important for intracellular signaling [10, 32, 42, 43]. No comprehensive studies of acetylation and methylation have been reported, but the availability of specific antibodies to enrich acetylated or methylated proteins opens up a possibility of such research [42, 43]. Most commonly, acetylation and methylation have been studied in selected proteins, and aimed at identification of sites of modifications. Although it is of a limited value for modeling, just description of PTM sites in a single protein contributes to the building of a network of functional relations between proteins.

Multiplexed use of fluorescence probes recognizing specific PTMs is an interesting approach to the comprehensive analysis of protein phosphorylation, glycosylation and expression [44]. It is especially attractive, as this technique allows analysis of two different PTMs in the same gel using probes that detect various PTMs. Many proteins have various PTMs simultaneously, with one PTM affecting another modification. As an example, palmitoylation of Ras protein enhances an efficiency of Ras phosphorylation upon activation of tyrosine kinase receptors [45]. As another example, phosphorylation of Smad3 on the C-terminal serine residues has been shown to promote ubiquitination of Smad3 [46]. Such examples suggest that combinations of PTMs are important for definition of functional activities of proteins, and these combinations have to be considered in modeling. Therefore, the development of techniques for simultaneous studying various PTMs is of great importance, and has to be intensified.

Proteins execute their functions in transient or stable complexes with other molecules, and descriptions of such complexes indicate functional features of proteins [3, 47, 48]. Purification of protein complexes, followed by an identification of their components, has been performed on a large scale in a number of studies. Thousands of interactions have been described when a large number of proteins were used as baits [49, 50], and in studies with selected signaling-related bait proteins [51–55]. For example, 131 proteins were identified in the TNFα-dependent signaling by a protein-protein interaction screen with 32 baits; many of these proteins indicated novel functions of TNFα [51]. Novel regulatory mechanisms in bone morphogenetic proteins and TGFβ signaling have been suggested by the 33 and 26 proteins found in complexes with BMPR-II and Smad3, respectively [54, 55]. These and other studies have contributed to the definition of connections between protein species, which are essential for design of model architecture.

Two-hybrid and phage display techniques also contribute to building protein-protein interaction networks [56–60]. The two-hybrid screen in *D. melanogaster* identified 4780 interactions involving 4679 proteins [57]. The yeast two-hybrid (Y2H) screen with *C. elegans* proteins detected 5500 interactions [58], and a number of novel interactions have been discovered in TGFβ signaling by yeast two-hybrid assays [59]. Studies of proteins binding to a particular peptide motive have contributed to the characterization of functional relations between proteins [60].

Comparison of datasets obtained with various techniques has shown that these datasets complement each other, and none of the techniques has provided full coverage of protein-protein interactions. As an example, Y2H assays would identify mostly binary interactions, while pull-down assays would describe predominantly protein complexes [47, 48, 61]. Localization of interactions is also of importance, since, for detection by Y2H technique, proteins have to be translocated to the nucleus [47, 48]. In pull-down assays, intracellular localization of the bait and interacting proteins may affect which interactions will be detected. For instance, to identify interacting partners of a nuclear protein, the bait-protein has to be localized in the nucleus. Protein-protein interactions also have various affinities, and can be stable or transient [47, 48, 61]. Some techniques allow detection of transient interactions, *e.g.*, Y2H, while other will detect only stable complexes. The last can be exemplified by the tandem-affinity purification, which detects only complexes which are sufficiently stable to be preserved after two sequential precipitation procedures [49]. Thus, considerations of binary interactions *versus* protein complex detection, localization of proteins, and differences in affinities of interactions strongly require combination of protein-protein interaction data obtained by different techniques. Such combined databases would allow validation of interactions across various datasets obtained in various cells, tissues and species.

An important issue for building of protein-protein interaction networks is often poor quality of data. As an example, in some two-hybrid screens, the level of false-negative and false-positive interactions can be up to 50% [3, 48, 56–61]. Confirmation of an interaction by different techniques, and identification of the same interaction in separate projects are expected to alleviate the problem of recording of false data. [3, 48, 56–61].

Proteomes of every cell type are unique. This suggests an importance of selection of a model system for studying protein-protein interactions. As an example, protein-protein interactions identified in yeast may not be observed in mammalian cells, and *vice versa*. This reflects cooperative nature of protein-protein interactions, when strength of an interaction is dependent on proteins PTMs, exposure of protein surfaces upon synthesis and folding of proteins, and milieu of the interaction, *e.g.*, local intracellular pH, osmolarity and ionic strength [47, 48, 61]. As conditions for interactions, folding and PTMs pattern of the same protein may differ in various cells, acquisition of protein-protein interaction data has to consider the model used, *e.g.*, species (mammalian, yeast, worm, fly or frog cells) and histological origin (epithelial, mesenchymal or other).

Protein complexes may interact with each other. Knowledge of complexes that functionally interact via the exchange of identified proteins may lead to the understanding of the functional importance of these proteins. It would define the hubs and connections in functional networks. Gagneur *et al.* [62] presented an example of modular decomposition for a set of data which described complexes formed in TNFα sig-

naling. Identification of series of interactions within one complex or module, and parallel connections of a protein to different complexes in protein-protein interaction networks defined the dynamics of complexes, and unveiled proteins important for the architecture of distribution and the multiplication of signaling [62]. This approach proves that modeling tools can be applied to protein-protein interaction networks to describe their dynamics.

An analysis of intracellular organelles is another way to approach protein-protein interactions. A number of protein complexes have been described in mitochondria (complex I, V, mitochondrial ribosomal complex), in chloroplast (photosystems I and II antenna proteins, chloroplast ribosomal complexes), in Golgi substructures (peripheral and integral membranes), in lipid rafts (as parts of the plasma membrane where signaling activity is concentrated), in the nucleus (RNA polymerase complexes, PolII; nucleolus, nuclear-pore complexes, PolII pre-initiation complex), and in the structure called midbody, which is a spindle midzone structure containing proteins indispensable for cytokinesis, asymmetric cells division and chromosome segregation [63–69]. The definition of protein composition of organelles does not indicate whether interactions between proteins are direct. This uncertainty of the protein-protein connections in organelles is a problem for the modeling of networks of protein interactions. On the other hand, the composition of organelles provides valuable information about functional complexes and their regulation, and therefore contributes to the definition of functional clusters in models.

Direct assessment of protein activities has been approached by activity-based proteome profiling (ABPP), which is based on the selection of proteins with a particular activity [18, 70]. As an example, an exploration of active serine proteases in a comprehensive way was performed by using chemical probes which interacted with and labeled only active enzymes [70]. The detection of substrates of proteases and kinases, and comprehensive analysis of ATP-binding proteins have been reported [71, 72]. Modification of the ATP-binding pocket in kinases in such a way that it can recognize only modified ATP, and has significantly decreased affinity to natural non-modified ATP, allows monitoring of substrates of the selected kinase in intact cells [73]. The last technique has not yet been used for a large-scale proteomics-based search, but the possibility of simultaneously monitoring all substrates of a kinase in an intact cell is of thrilling interest for modeling. Performed over a time interval, this technique may provide a tool for evaluation of kinetic parameters of substrate phosphorylation.

The available proteomics tools for studying PTMs, protein-protein interactions and the evaluation of protein activities are capable of generating large volumes of data (Fig. 4). The comprehensiveness and preciseness of these datasets vary, but in many cases they fulfill minimal requirements for being suitable for modeling, *e.g.*, requirement for description of a number of components and relation between them, which is sufficient for building a model. Notably, proteomics datasets describe protein species in sufficient numbers, and explore dependencies between these protein species in the network. Further standardization of data generation, as well as insights into the dynamics of the functional status of proteins will significantly increase the value of proteomics data. This is the main challenge for proteomics today, and it requires further technical developments of a real-time mode proteomics.

1.5 Localization of proteins

Modeling of intracellular reactions with the assumption of cells as uniform chemical reactors is often incorrect. To avoid this mistake, descriptions of protein localization inside of cells have to be provided. Protein localization has to be described in intracellular compartments, *e.g.*, the nucleus or cytoplasm, and also in organelles, as specialization of cellular organelles defines the functional roles of proteins. The most informative would be data about protein localization in an intact cell, although it is challenging technically.

Various microscopy-based techniques have been employed to monitor green fluorescent protein (GFP)- or yellow fluorescence protein (YFP)-tagged proteins in yeast and human cells [74–76]. The most comprehensive work has been performed with yeast proteins, with definition of 22 categories for subcellular localizations of 75% of yeast proteins [74], and the description of the subcellular localization of 6100 proteins [77]. Co-localization of selected proteins can also be monitored in an intact cell by fluorescence resonance energy transfer technique (FRET) [78]. Co-localization of signaling proteins has been known as an important condition for efficient signaling [3, 79]. Stochastic models of protein-protein interaction have shown that co-localization is a feature of functional complexes of proteins, which do not necessarily form physical complexes [80]. For proteomics studies, it indicates a requirement for the description of protein localization even within a single intracellular compartment, *e.g.*, cytoplasm or the nucleus. Thus, a comprehensive definition of protein localization can be achieved by a combination of microscopy-based techniques which describe localization of proteins, with proteomics studies which describe functional features of proteins (Fig. 5).

Studies of the protein content of organelles, as a way to unveil protein-protein interactions, have been discussed above (see Section 1.3). Organelle-specific localization of proteins is also required for the modeling of functional relations between protein species. Resources which predict the organelle-specific localization of proteins have been described, *e.g.*, MitoProt, ChloroP, PredictNLS, SignalP, iPSORT, TagetT (reviewed in [63]). The drawback of many reported proteomics studies of organelles is that they do not consider organelle-specific proteins which are not incorporated into organelles, as only an organelle-bound part of the total pool of selected protein was studied. This diminishes the value of such a study for modeling purposes.

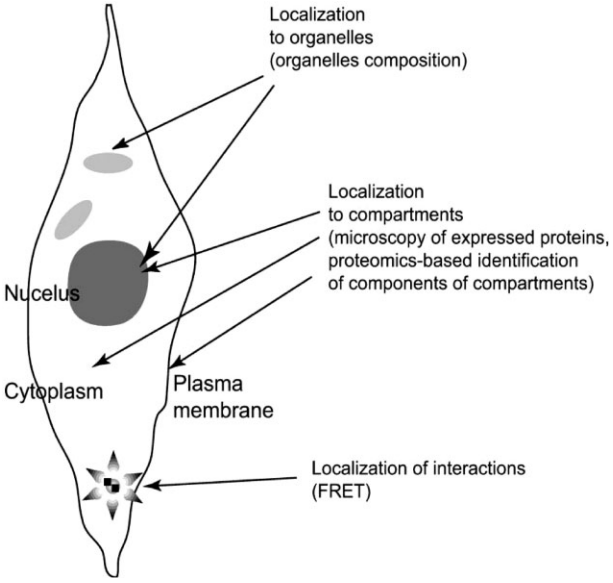

Figure 5. Description of protein localization. Intracellular localization of proteins can be evaluated by microscopy, protein-protein interaction techniques, *e.g.*, FRET, and the identification of proteins in various organelles and cellular structures, *e.g.*, the cytoplasm, nucleus, plasma membrane, mitochondria, etc. Identities of studied proteins can be provided by 2-DE/MS, LC-MS/MS, and by use of protein expression libraries.

Important sources for validation of the localization of selected proteins are databases of life-science publications, *e.g.*, PubMed. These databases may provide information about the intracellular distribution of protein species, which is complementary to high-throughput approaches using proteomics. Currently, significant human resources are dedicated to evaluating the quality of published data: researchers have to read every paper before deciding whether reported data are relevant to the aim of a study [81]. This is obviously inefficient, and it urges for the development of search engines capable of the unsupervised extraction of information about protein localization and functions. Such search engines may significantly simplify selection of papers which will be further evaluated by researchers.

In modeling, the localization of proteins defines compartments for the calculation of functional dependencies. It is important that the distribution of proteins in various compartments is measured, not only in a single given compartment. The dynamics of protein distribution between various compartments and organelles allow for the definition of constants for the diffusion and active and passive transport of protein species. Description of protein localization would require use of a non-destructive technique, *e.g.*, microscopy, combined with destructive 2-DE, LC-MS/MS and other high-throughput proteomics and protein expression techniques (Fig. 5). It is essential that the representation of data generated by the various approaches will be in the same modeling-compatible format.

1.6 Dynamics of proteomes

The time scale of biological processes involving proteins vary from milliseconds to days. Knowledge of a time scale of studied processes is essential for modeling. It defines an interval of time and frequency of data collection, which are required for capturing features of a modeled process. The shortest time scale is characteristic for chemical reactions with proteins; PTMs and protein-protein interactions occur in milliseconds to seconds [82–84]. An accumulation of modified proteins may take a longer time, although accumulation rates are dependent on parameters of fast reactions with single proteins. This suggests that the time frame of fast reactions with single protein molecules is the first factor, which defines the dynamics of proteome changes.

Regulatory processes include multiple reactions between hundreds of individual proteins. The time frames of such processes may vary from milliseconds to hours. Examples of processes with a short lag period are electro-physiological reactions: it takes milliseconds to generate an active potential in neurons [82]. The intracellular signaling process initiated by growth factors may take from minutes to hours [3, 79]. The longer time period, as compared to single-protein modifications, is the result of re-arrangements of interactions between proteins, and movements of proteins [28, 83–90]. Signaling processes consist of sequential and time-parallel changes of various protein species. Every step in such a signaling process requires time for transformation of protein species, *e.g.*, by PTMs, and for rearrangements of the protein-protein interactions [91]. The architecture of connections between protein species was found to be crucial for the formation of thresholds, oscillations and bi-stabilities [2, 90, 91]. Thus, a number of steps with various time dependencies in regulatory cascades is the second factor that influences the dynamics of regulatory processes.

Protein movements constitute the third factor that defines the dynamic features of proteomes. Substrates and components of complexes have to reach intracellular sites where reactions of modifications and complex formation take place, and they have to be transported to sites of further signal propagation. Estimated diffusion coefficients for most of biologically active proteins are 0.1–1.0 $\mu m^2/s$ [75]. It means that a protein may be homogenously distributed within a cell in less than 10 min [75]. However, cells are not homogenous chemical reactors. Compartmentalization of cells creates structural barriers for free protein diffusion, and the activities of transport systems affect the duration of signaling processes by introducing time delays, from seconds up to hours [75, 76, 91–94]. Many diffusion and transport coefficients have been measured in directed studies with a limited number of proteins. The lack of comprehensive large-scale studies of protein transport systems has been hampering an efficient employment of distribution coefficients in modeling efforts, and is urging development of high-throughput approaches. Thus, the kinetic of reactions with single protein molecules, the number of reactions and

connections between protein species, and the movements of proteins constitute three factors which define dynamic properties of proteomes (Fig. 6).

Most of the proteomics experiments with time frames longer than 24 h explore the long-term transformations of cells [95, 96]. These effects may be related to cells of a selected type, *e.g.*, exploring the differentiation of cells, or to cells in a context of multicellular organisms, *e.g.*, the study of cells in a developing body. This includes the profiling of various tissues and different cell lines that originate from the same tissue [95, 96]. Thus, longer time frames in proteomics studies are associated with the larger complexity of the studied system (Fig. 6). The same paradigm is valid for modeling: large systems with multiple time-delaying connections require analysis over longer time periods, as compared to simple systems with a small number of variables [2, 8, 83, 97–99].

Proteomes are complex systems with thousands of connected protein species. Thus, for the most complete datasets, researchers would have to collect data about single molecules with a time interval of a few milliseconds over a period of days. This is obviously not a feasible task, as techniques for studying single molecules are not suitable for large-scale projects, and proteomics is still too laborious to be able to analyze hundreds of experimental time points in a single project [5, 6]. The modeling-based definition of protein complexes and functional modules [100, 101] has provided a solution to such enormous tasks by dividing them into feasible parts, and then integrating generated datasets. For example, separate studies of protein complexes, organelles, and signaling pathways can be integrated in one model. Current modeling tools allow for doing that, under the condition of unified data acquisition and representation.

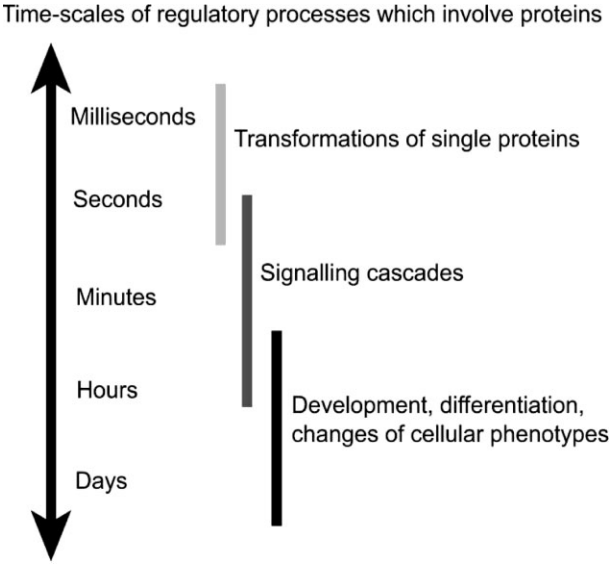

Figure 6. Time scale for various regulatory processes involving proteins is from milliseconds to days. Transformations of single protein molecules, *e.g.*, PTMs, represent fastest processes, signaling cascades may take from seconds to hours, and changes of cellular phenotypes, *e.g.*, differentiation, represent processes which may take days.

An important issue in study of proteome dynamics is a definition of time intervals and a number of points for data collection. In most cases, the definition of frequency and time-intervals for data collection are based on an empirical knowledge of studied systems, *e.g.*, a knowledge posterior to modeling [28, 79, 83–89]. Notably, it is based on knowledge of how long a time it takes to observe changes of interest, and on knowledge of the dynamics of changes, *e.g.*, features of cell division, cell death, differentiation, etc. Balanced with a workload, it defines a number of points for data collection. Such definition is prone to false conclusions, as it is based on subjective assumptions instead of calculated predictions. Solving the problem of optimization of data collection intervals would significantly improve the design of experiments. An example of such an approach has been reported by Wolkenhauser and colleagues [102], who applied a multiple-shooting method to a system of ordinary differential equations describing a simple dynamic model. This concept of time-point selection has proven to enhance confidence in experimental data at least fivefold. A possibility of calculating the frequency and number of sampling points would strongly enhance the efficiency of proteomics studies, as it will warrant collection of data at time points significant for modeling, and will capture all the crucial dynamic features of a proteome. The concept for the optimization of sampling time has been developed for simple dynamic systems, and its application to much more complex systems, *e.g.*, proteomes, is the next challenge.

1.7 From signaling pathways to a network signaling via proteomics and systems biology

Recent progress in cell biology indicates that the concept of signaling pathways is to be substituted by a concept of network signaling. The first feature of the network signaling concept coins multiplicity and inter-dependence of various outputs in response to one input, while the signaling pathway concept indicates linear distribution of a signal from one input to various output with limited or even unconsidered dependencies (Fig. 7A, B). The second specific feature of network signaling, as compared to signaling pathways, is the consideration of a single variable input together with a number of non-variable inputs, which modulate the variable one and can change the values of their effects depending on the dynamics of the studied input (Fig. 7B). The signaling pathways concept generally disregards the values of non-variable inputs, and it mentions them only in cases of non-interpretable results (Fig. 7A). In cell biology, such difficult-to-interpret results are often presented as being influenced by a cell-type-specific background.

As an example, studies of TGFβ signaling resulted in a relative linear pathway, which contains ligand, cell surface receptors, intracellular Smad-dependent and Smad-inde-

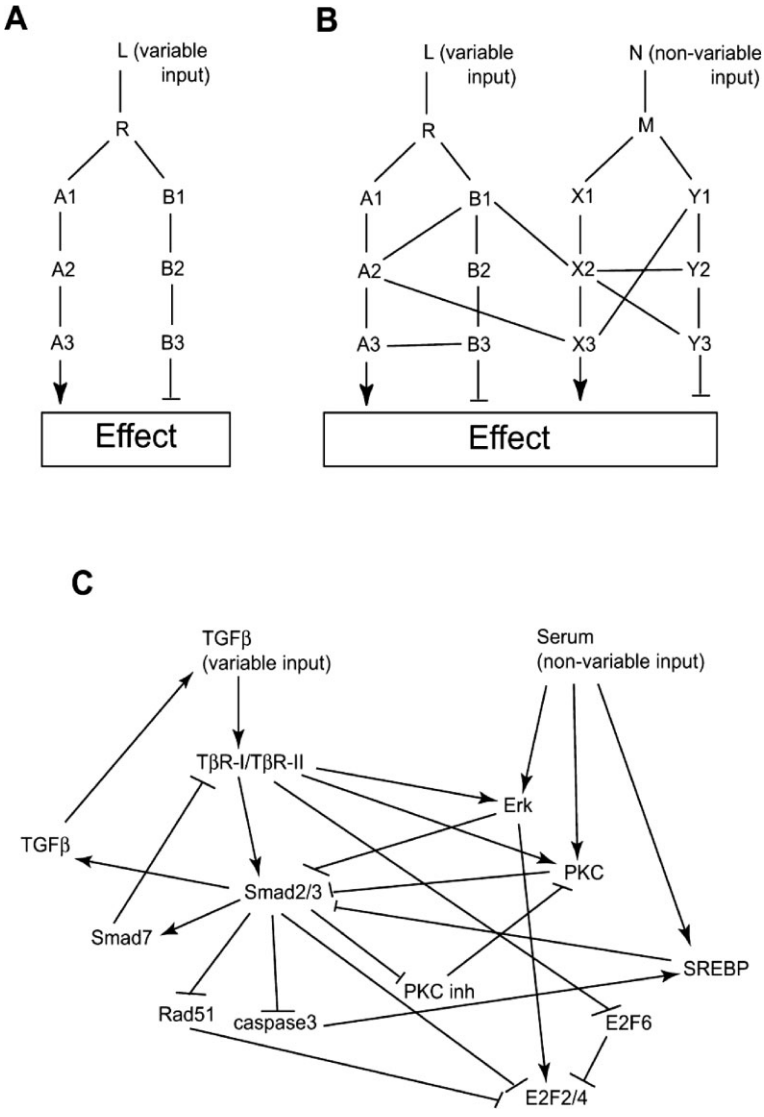

Figure 7. Network signaling concept, as an alternative to signaling pathways concept. (A) Representation of signaling pathways, as pathways which are dependent exclusively on a variable input value (L). Output effect for such a model depends on the intensities of input-initiated pathways (A1-3 and B1-3). (B) Network signaling model is represented by interconnected pathways, which are dependent on variable input (L), as well as on non-variable input (N) values. The output effect in such a model is dependent on both variable and non-variable inputs, as there are extensive dependencies between protein species. L, N, R, A(1-3), B(1-3), X(1-3), and Y(1-3) represent components of models, *e.g.*, protein species. (C) TGFβ signaling is shown as a network of protein species. Variable input is represented by TGFβ, and non-variable input is represented by serum. TβR-I/TβR-II, Smad2/3, Smad7, Erk, E2F6, E2F2/4, Rad51, PKC, PKC inhibitor, SREBP and caspase-3 are proteins regulated by TGFβ. Direct effects of the serum on Erk, PKC and SREBP are indicated. Note a number of feedforward and feedback loops in the network. Proteins shown represent only a part of proteins known to be regulated by TGFβ. Stimulatory effects are indicated by arrows, and inhibitory effects are shown by T-lines (A–C).

pendent pathways, leading to regulation of gene expression [12, 34]. However, a number of publications have shown that this simplified concept is incorrect, especially in the context of biological processes [12, 34, 54, 55, 88]. A number of signal amplification and redistribution points with a variable architecture of connections have been reported: *e.g.*, receptors, Smad proteins, Erk and PKC kinases (Fig. 7C). The described connections between some of the proteins involved in TGFβ signaling form a number of feedback and feed-forward loops, as shown in Fig. 7C. In addition, an inability to explain results obtained in studies of *in vivo* systems, which were not manipulated by overexpression or down-regulation of selected components (reviewed in [34]), is a strong indication that the concept of TGFβ signaling pathway has to be revised. TGFβ has been described as a potent inhibitor of cell proliferation. However, cell culture conditions (non-variable inputs) may change the efficiency of inhibition of cell proliferation from 90% of inhibition in the presence of 3% of serum in culture medium, to the maximum 50%, in the presence of 10% serum in medium (our unpublished observation, Mv1Lu cells; Fig. 7C). Thus, for the prediction of results many more factors have to be considered than concentration of the ligand and the presence of specific receptors. Attempts to understand signaling networks by studying their fragments are doomed to failure. An informative approach requires a systematic study of a significant number of variables which represent a signaling network, and not only a pathway. Traditional biological experiments can seldom generate information in quantities suitable for modeling, while high-throughput techniques, *e.g.*, transcriptomics and proteomics, have the capacity to produce sufficiently rich datasets. However, the quality and format of such data still have to be adapted to the requirements of systems biology, *e.g.*, requirement of information about proteins quantities, functional status, localization and dynamics of proteome changes, as is discussed in this review.

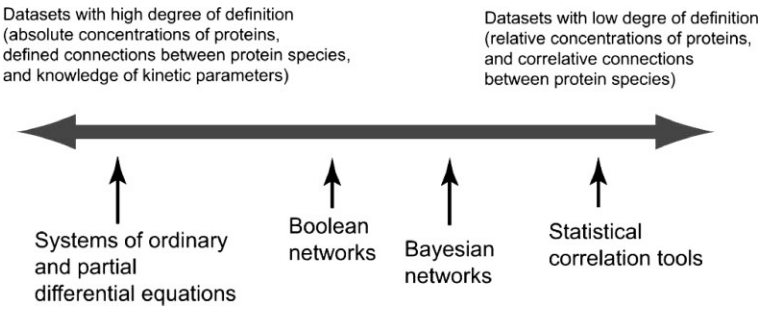

Figure 8. Application of modeling tools to datasets with various degree of definition. Proteomics produces data of various degrees of definition of protein species, *e.g.,* various levels of details about absolute or relative concentrations of proteins, their activities, localization and dynamics of changes of these proteins. Each set of such data can be most efficiently analyzed by modeling tools designed to process data with various definitions of details in the description of proteomes, as indicated. As examples, differential equations, Boolean and Bayesian networks, and statistical correlation tools are indicated. They are shown in relation to the required level of details about datasets to be efficiently analyzed with those tools.

On the other side of the proteomics-systems biology integration bridge, systems biology offers tools which can process results already available now. Modeling tools that can accommodate data of various degree of definition of details about components have been described [4, 7, 8, 97] (Fig. 8). High-degree definition models are based on a precise knowledge of features of datasets that are sufficiently rich to build a system of defined dependencies between variables, and of precisely calculated values of variables in dynamics. In mathematical terms, such models are based on differential equations. As the use of differential equations requires detailed input, such modeling has been successfully applied to functional modules, but not to cell-wide datasets. The modeling of signaling by Smad proteins [93], EGF receptors [83], heteromeric G proteins [89], TNF/NF-κB [103], are examples of this approach. Common to all these models is the requirement for information about concentrations and kinetic parameters of all involved variables (protein species). Lack of such information may lead to collapse of the model or to incorrect conclusions.

To overcome the requirement of high definition, modeling tools based on correlative relations have been developed. For instance, Boolean networks can be built with data about connections and the functional status of variables of a system. They also consider the number of inputs and outputs for components. Boolean networks can analyze data that describe the state of a single variable and the average state of all other neighbors [104–106]. They do not require knowledge about concentrations and the kinetic parameters of components. Bayesian networks allow analysis of even less well-defined data, as they describe statistical relationships between variables from a dataset, and do not require knowledge of confirmed connections between species. These statistical relationships may include pair-wise relations and relations between arbitrary complexes. Bayesian networks summarize dependencies between variables, and select for the highest probable model, given the data [107]. Essentially, Boolean and Bayesian networks are based on the statistical values of relations, which may be exemplified by the statement "if A and B are up, then C should be down". For proteomics datasets, it provides an opportunity to search for proteins with correlative behavior, *e.g.*, expression, pattern of phosphorylation, etc. (Fig. 6).

A low level of data definition may have a drawback of selecting models that may be correct only in a limited diapason of values of variables. As an example, an analysis of activation of FAK and MAPK resulted in two different models, because one model was built on data about the initial activation of FAK and MAPK, and the other on the analysis of steady-state activation of these two kinases [107]. Therefore, knowledge of a studied system, *e.g.*, biological data, defined as posterior probability, is an important component for selection of a correct model. It introduces real experimental parameters, including their limiting values, and prevents a model from collapse.

Statistical analysis of various types has also been applied to datasets with different definition of details about components, increasing a number of modeling approaches [3, 4, 100, 101].

Modeling of biological systems is affected by modeling of abstract complex systems. Modeling of abstract systems suggests the most optimal size, relations and dynamics of a system, which are conditions for acquiring such features as robustness and sensitivity. Such abstract mathematical modeling defines restriction rules for the modeling of biological systems: for instance, it describes how stability of a system will be affected by a number of affected variables, and by a number of inputs and outputs of these variables [104, 105, 108, 109]. Moreover, it also predicts the requirement of a certain level of noise, as random and non-specific changes in

dependencies and values, in order to have a system evolving [104]. Knowledge of the behavior of abstract mathematic models contributes to the establishment of rules for biologically relevant systems, which sets critical limits for quantity and features of variables, and improves the computing of biological models.

There are a number of studies of complex biological processes for which combined proteomics and large-scale modeling have been reported. As an example, the ICAT-based evaluation of protein expression contributed to identification of crucial interactions in metabolic networks, *e.g.*, the galactose-utilization pathway [98]. Correlation analysis of proteomics data describing interferon signaling in liver cells unveiled a number of novel components of the interferon response [99]. Constraint-based analysis of the mitochondrial metabolic network defined the optimal distribution of reaction flux [94]. These and other studies provide proofs that the combination of proteomics and modeling tools can generate novel knowledge. This knowledge may be essential for the design of efficient treatments of diseases, *e.g.*, understanding of the robustness and sensitivity of cancer cells may indicate proteins which, if being targeted, may promote regression of tumors [2].

Systems biology develops tools which make possible combination of datasets generated by various techniques, such as transcriptomics, proteomics and metabolomics data [3, 4, 82, 110–113]. These techniques have developed tools for acquisition and presentation of data in an interchangeable format, *e.g.*, XML-supporting MIAME, PEDRo, MAGE-ML. Further successful solutions for the quantification, functional, spatial and dynamic characterization of proteomes are essential for the integration of proteomics into the whole-cell modeling.

2 Concluding remarks

The integration of proteomics and large-scale modeling is required for the description of complex biological systems. Methods for such integration are now under development, and propose suitable designs for proteomics experiments. Here is one such "recipe". First, proteomics experiments describe a number of proteins of interest for a selected biological process. These protein species are annotated using dedicated software and are represented in XML format, with quantitative descriptions of PTMs and interacting partners. Information about the localization of proteins is recorded, and all data are collected in real-time dynamics. This dataset will be used to build a model to unveil dependencies between protein species. Then, the model describing these dependencies will be interrogated in biological experiments with the modulation of expression and activities of identified key proteins. This will validate the architecture of the model, and will define the levels of robustness and sensitivity of the studied biological process. Knowledge of crucial components of a system will lead to development of drugs that will target these components to achieve high efficiency of treatment with negligible side effects.

How many of such "recipes" are in a "cookbook" of biology? An answer to this question may be already available in the next few years, as progress in proteomics and systems biology brings these two fields together.

I apologize for not being able to cite all relevant works due to limitations of space. I am grateful to Hanna Woksepp and Anna Dubrovska for comments, and to Kate Williams for the help with manuscript preparation. This work was supported in part by grants from the Swedish Cancer Society, the Swedish Research Council, Hiroshima University and Merck KGaA.

3 References

[1] Butcher, E. C., Berg, E. L., Kunkel, E. J., *Nat. Biotechnol.* 2004, *22*, 1253–1259.

[2] Kitano, H., *Nat. Rev.* 2004, *4*, 227–235.

[3] Papin, J. A., Hunter, T., Palsson, B. O., Subramaniam, S., *Nat. Rev. Mol. Cell Biol.* 2005, *6*, 99–111.

[4] Ideker, T., Lauffenburger, D., *Trends Biotechnol.* 2003, *21*, 255–262.

[5] de Hoog, C., Mann, M., *Annu. Rev. Genomics Hum. Genet.* 2004, *5*, 267–293.

[6] Görg, A., Weiss, W., Dunn, M. J., *Proteomics* 2004, *4*, 3665–3685.

[7] Aggarwal, K., Lee, K. H., *Brief. Funct. Genomics Proteomics* 2003, *2*, 175–184.

[8] Cho, K.-H., Wolkenhauer, O., *Biochem. Soc. Trans.* 2003, *31*, 1503–1509.

[9] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H. *et al.*, *Bioinformatics* 2003, *19*, 524–531.

[10] Yang, X.-J., *Oncogene* 2005, *24*, 1653–1662.

[11] Pawson, T., *Cell* 2004, *116*, 191–203.

[12] Heldin, C.-H., Purton, M., *Signal transduction*, Chapman and Hall, London 1996, pp. 1–365.

[13] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J. *et al.*, *Nat. Biotechnol.* 2004, *22*, 177–183.

[14] Taylor, C. F., Paton, N. W., Garwood, K. L., Kirby, P. D. *et al.*, *Nat. Biotechnol.* 2003, *21*, 247–254.

[15] Pedrioli, P. G. A., Eng, J. K., Hubley, R., Vogelzang, M. *et al.*, *Nat. Biotechnol.* 2004, *22*, 1459–1466.

[16] Stanislaus, R., Jiang, L. H., Swartz, M., Arthur, J., Almeida, J. S., *BMC Bioinformatics* 2004, *5*, 9: 1–7.

[17] http://www.geneontology.org

[18] Adam, G. C., Sorensen, E. J., Cravatt, B. F., *Mol. Cell. Proteomics* 2002, *1*, 781–790.

[19] Cutler, P., *Proteomics* 2003, *3*, 3–18.

[20] Oda, Y., Nagasu, T., Chait, B. T., *Nat. Biotechnol.* 2001, *19*, 379–382.

[21] Zhou, H., Watts, J. D., Aebersold, R., *Nat. Biotechnol.* 2001, *19*, 375–378.

[22] Vosseller, K., Hansen, K. C., Chalkley, R. J., Trinidad, J. C. *et al.*, *Proteomics* 2005, *5*, 388–398.

[23] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F. *et al.*, *Nat. Biotechnol.* 1999, *17*, 994–999.

[24] Ong, S. E., Foster, L. J., Mann, M., *Methods* 2003, *29*, 124–130.

[25] Julka, S., Regnier, F., *J. Proteome Res.* 2004, *3*, 350–363.

[26] Schmidt, A., Kellermann, J., Lottspeich, F., *Proteomics* 2005, *5*, 4–15.

[27] Gevaert, K., Goethals, M., Martens, L., Van Damme, J. *et al.*, *Nat. Biothecnol.* 2003, *21*, 566–569.

[28] Blagoev, B., Ong, S.-E., Kratchmarova, I., Mann, M., *Nat. Biotechnol.* 2004, *22*, 1139–1145.

[29] Reinders, J., Lewandowski, U., Moebius, J., Wagner, Y., Sickmann, A., *Proteomics* 2004, *4*, 3786–3703.

[30] Traxler, E., Bayer, E., Stöckl, J., Mohr, T. *et al.*, *Proteomics* 2004, *4*, 1314–1323.

[31] Gerner, C., Vejda, S., Gelbmann, D., Bayer, E. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 528–537.

[32] Greighton, T. E., *Proteins: Structure and molecular properties*, Freeman, New York1993, pp. 78–102.

[33] Pawson, T., *Cell* 2004, *116*, 191–203.

[34] Souchelnytskyi, S., *Exp. Oncol.* 2002, *24*, 3–12.

[35] Dubrovska, A., Souchelnytskyi, S., *Proteomics* 2005, in press.

[36] Guy, G. R., Philip, R., Tan, Y. H., *Electrophoresis* 1994, *15*, 417–440.

[37] Stasyk, T., Dubrovska, A., Lomnytska, M., Yakymovych, I. *et al.*, *Mol. Biol. Cell* 2005, DOI: 10.1091/mbc.E05-03-0257.

[38] Knight, Z. A., Schilling, B., Row, R. H., Kenski, D. M. *et al.*, *Nat. Biotechnol.* 2003, *21*, 1047–1054.

[39] Hirabayashi, J., Kasai, K.-I., *J. Chromatogr.* 2002, *721*, 67–87.

[40] Peng, J., Schwartz, D., Elias, J. E., Thoreen, C. C. *et al.*, *Nat. Biotechnol.* 2004, *21*, 921–926.

[41] Vertegaal, A. C. O., Ogg, S. C., Jaffray, E., Rodriguez, M. S. *et al.*, *J. Biol. Chem.* 2004, *279*, 33791–33798.

[42] Fu, M., Wang, C., Wang, J., Zafonte, B. T. *et al.*, *Cytokine Growth Factor Rev.* 2002, *13*, 259–276.

[43] Umlauf, D., Goto, Y., Feil, R., *Methods Mol Biol.* 2004, *287*, 99–120.

[44] Ge, Y., Rajkumar, L., Guzman, R. C., Nandi, S. *et al.*, *Proteomics* 2004, *4*, 3464–3467.

[45] Dudler, T., Gelb, M. H., *J. Biol. Chem.* 1996, *271*, 11541–11547.

[46] Fukuchi, M., Imamura, T., Chiba, T., Ebisawa, T. *et al.*, *Mol. Biol. Cell.* 2001, *12*, 1431–1443.

[47] Jensen, L. J., Bork, P., *DDT:Targets* 2004, *3*, 51–56.

[48] Bader, G. D., Hogue, C. W. V., *Nat. Biotechnol.* 2002, *20*, 991–997.

[49] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[50] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[51] Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.-O. *et al.*, *Nat. Cell Biol.* 2004, *6*, 97–105.

[52] Tran, H. T., Ulke, A., Morrice, N., Johannes, C. J., Moorhead, G. B. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 257–265.

[53] Zhou, Y., Gu, G., Goodlett, D. R., Zhang, T. *et al.*, *J Biol. Chem.* 2004, *279*, 39155–39164.

[54] Hassel, S., Eichner, A., Yakymovych, M., Hellman, U. *et al.*, *Proteomics* 2004, *4*, 1346–1358.

[55] Grimsby, S., Jaensson, H., Dubrovska, A., Lomnytska, M. *et al.*, *FEBS Lett.* 2004, *577*, 93–100.

[56] Phizicky, E. M., Fields, S., *Microbiol. Rev.* 1999, *59*, 94–123.

[57] Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A. *et al.*, *Science* 2003, *302*, 1727–1736.

[58] Li, S., Armstrong, C. M., Bertin, N., Ge, H. *et al.*, *Science* 2003, *303*, 540–543.

[59] Colland, F., Jacq, X., Trouplin, V., Mougin, C. *et al.*, *Genome Res.* 2004, *14*, 1324–1332.

[60] Elia, A. E. H., Cantley, L. C., Yaffe, M. B., *Science* 2003, *299*, 1228–1231.

[61] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2002, *417*, 399–403.

[62] Gagneur, J., Krause, R., Bouwmeester, T., Casari, G., *Genome Biol.* 2004, *5*, R57.

[63] Warnock, D. E., Fahy, E., Taylor, S. W., *Mass Spectrom. Rev.* 2004, *23*, 259–280.

[64] Skop, A. R., Liu, H., Yates, J. III, Meyer, B. J., Heald, R., *Science* 2004, *305*, 61–66.

[65] Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P. *et al.*, *Nature* 2003, *426*, 570–574.

[66] Sickmann, A., Reinders, J., Wagner, Y., Joppich, C. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 13207–13212.

[67] Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M. *et al.*, *Nat. Biotechnol.* 2003, *21*, 281–286.

[68] Mootha, V. K., Bunkenborg, J., Olsen, J., V., Hjerrild, M. *et al.*, *Cell* 2003, *115*, 629–640.

[69] Brunet, S., Thibault, P., Gagnon, E., Kearney, P. *et al.*, *Trends Cell Biol.* 2003, *13*, 629–638.

[70] Saghatelian, A., Jessani, N., Joseph, A., Humphrey, M., Cravatt, B. F., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 10000–10005.

[71] Bredmeyer, A. J., Lweis, R. M., Malone, J. P., Davis, A. E. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 11785–11790.

[72] Besant, P. G., Lasker, M. V., Bui, C. D., Tan, E. *et al.*, *J. Proteome Res.* 2004, *3*, 120–125.

[73] Liu, Y., Shah, K., Yang, F., Witucki, L., Shokat, K. M., *Chem. Biol.* 1998, *5*, 91–101.

[74] Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S. *et al.*, *Nature* 2003, *425*, 686–691.

[75] Roix, J., Mistelli, T., *Histochem. Cell. Biol.* 2002, *118*, 105–116.

[76] Liebel, U., Starkuviene, V., Erfle, H., Simpson, J. C. *et al.*, *FEBS Lett.* 2003, *554*, 394–398.

[77] Kumar, A., Agarwal, S., Heyman, J. A., Matson, S. *et al.*, *Genes Dev.* 2002, *16*, 707–719.

[78] Mahajan, N. P., Linder, K., Berry, G., Gordon, G. W., *et al*,. *Nat. Biotechnol.* 1998, *16*, 547–552.

[79] Hlavacek, W. S., Faeder, J. R., Blinov, M. L., Perelson, A. S., Goldstein, B., *Biotechnol. Bioeng.* 2003, *84*, 783–794.

[80] Batada, N. N., Shepp, L. A., Siegmund, D. O., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 6445–6449.

[81] Peri, S., Navarro, D., Amanchy, R., Kristiansen, T. Z. *et al.*, *Genome Res.* 2003, *13*, 2363–2371.

[82] Noble, D., *BioEssays* 2002, *24*, 1155–1163.

[83] Schoeberl, B., Eichler-Josson, C., Gilles, E. D., Muller, G., *Nat. Biotechnol.* 2002, *20*, 370–375.

[84] Ballif, B. A., Roux, P. P., Gerber, S. A., MacKeigan, J. P. *et al.*, *Proc. Natl. Acad. Sci. USA* 2005, *102*, 667–672.

[85] Pomerening, J. R., Sontag, E. D., Ferrell, J. E., *Nat. Cell Biol.* 2003, *5*, 346–351.

[86] Soskic, V., Godovac, M., Poznanovic, S., Boehmer, F. D., Godovac-Zimmermann, J., *Biochemistry* 1999, *38*, 1757–1764.

[87] El Yazidi-Belkoura, I., Adriaenssens, E., Dolle, L., Descamps, S., Hondermarck, H., *J. Biol. Chem.* 2003, *278*, 16952–16956.

[88] Kanamoto, T., Hellman, U., Heldin, C.-H., Souchelnytskyi, S., *EMBO J.* 2002, *21*, 1219–1230.

[89] Bornheimer, S. J., Maurya, M. R., Farquhar, M. G., Subramaniam, S., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 15899–15904.

[90] Ishii, N., Robert, M., Nakayama, Y., Kanai, A., Tomita, M. J., *Biotechnology* 2004, *113*, 281–294.

[91] Hasty, J., McMillen, D., Collins, J. J., *Nature* 2002, *420*, 224–230.

[92] Ferrigno, P., Silver, P., *Oncogene* 1999, *18*, 6129–6134.

[93] Kyoda, K. M., Muraki, M., Kitano, H., *Pac. Symp. Biocomput.* 2000, 317–328.

[94] Vo, T. D., Greenberg, H. J., Palsson, B. O., *J. Biol. Chem.* 2004, *279*, 39532–39540.

[95] Behre, G., Reddy, V. A., Tenen, D. G., Hiddemann, W., Zada, A. A., Singh, S. M., *Expert. Opin. Ther. Targets* 2002, *6*, 491–495.

[96] Souchelnytskyi, S., *J. Mammary Gland Biol. Neoplasia* 2002, *7*, 39–371.

[97] Eungdamrong, N. J., Iyengar, R., *Biol. Cell* 2004, *96*, 355–362.

[98] Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R. *et al.*, *Science* 2001, *292*, 929–933.

[99] Yan, W., Lee, H., Yi, E. C., Reiss, D. *et al.*, *Genome Biol.* 2004, *5*, R54.

[100] Plavec, I., Sirenko, O., Privat, S., Wang, Y. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 1223–1228.

[101] Spirin, V., Mirny, L. A., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 12123–12128.

[102] Kutalik, Z., Cho, K-H., Wolkenhauer, O., *BioSystems* 2004, *75*, 43–45.

[103] Cho, K.-H., Shin, S.-Y., Lee, H.-W., Wolkenhauer, O., *Genome Res.* 2003, *13*, 2413–2422.

[104] Amaral, L. A. N., Diaz-Guilera, A., Moreira, A. A., Goldberger, A. L., Lipsitz, L. A., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 15551–15555.

[105] Kauffman, S., Peterson, C., Samuelsson, B., Troein, C., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 17102–17107.

[106] Huang, S., Ingber, D. E., *Exp. Cell Res.* 2000, *261*, 91–103.

[107] Sachs, K., Gifford, D., Jaakkola, T., Sorger, P. Lauffenburger, D. A., *Science's STKE* 2002, http://stke.sciencemag.org/cgi/content/full/sigtrans;2002/148/pe38.

[108] Angeli, D., Ferrell, J. E., Sontag, E. D., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 1822–1827.

[109] Song, C., Havlin, S., Makse, H. A., *Nature* 2005, *433*, 392–395.

[110] Van der Greef, J., Stroobant, P., van der Heijden, R., *Curr. Opin Chem. Biol.* 2004, *8*, 559–565.

[111] Weston, A.D., Hood, L., *J. Prot. Res.* 2004, *3*, 179–196.

[112] Oksman-Caldentey, K-M., Inzé, D., Oresic, M., *Proc. Natl. Acad. Sci. USA* 2004, *101*, 9949–9950.

[113] Mayr, M., Mayr, U., Chung, Y.-L., Yin, X. *et al.*, *Proteomics* 2004, *4*, 3751–3761.

